On Biplots for *Principal Component Analysis* (PCA), *Between Class Analysis* (BCA) and *Canonical Variates Analysis* (CVA).

Gerhard Welzl

February 22, 2018

1 Introduction

Biplots are representations of multivariate data with information on both the samples and the variables of a data matrix. Placement of the points (samples) is directed by the used dimension reduction method (PCA, BCA, CVA or others). In particular PCA biplots are widespread. The common presentation of variables (usually by arrows) and samples allows both the interpretation of sample cluster and the correlation between variables; in addition relations between variables and objects can be described. PCA is based on the data matrix only, additionally informations about the samples are not utilized in *unsupervised* methods. However such factors (typically factors describing the experimental design) can be used for labelling samples in the biplot. A congruence of object clusters and factors is a strong hint for an association between variables and factors.

Biplots in connection with a *between class analysis* (BCA) are also popular, especially since a program `bca` in `ade4` package (R-project) is available. Under this name a *diagonal discriminant analysis* (DDA) is performed. This method is often used analyzing big data matrices with a high number of variables for only a small number of objects. The crucial aspect with this method is that no correlation matrix is estimated, assuming independent variables.

DDA is a special form within the scope of *supervised* methods. In general these methods are used to analyse a data matrix plus an additional *supervisor* with information about the objects (usually the experimental design). This information is used to construct a linear combination of all variables which allows a *best* separation of groups defined by the *supervisor*. This *Canonical Variates Analysis* (CVA) can be considered as a sequence of two PCAs; nevertheless biplots are less usual.

Despite the widespread use of biplot there a some discussions about the correct presentation and interpretation of biplot:

John C. Gower: Unified Biplot Geometry. Developments in Applied Statistics, 2003: *... some of the faults often found in published biplot diagrams:*

- *Unequal scaling of the axes*
- *use of E-formats*

- *ugly scale divisions*
- *no scales on biplot axes*
- *one-standard error lengths of vectors*
- *scales given unnecessarily for canonical axes and no scales for the original variables*
- *separate canonical scales for samples and for variables*

Some of these *faults* are also present using standard biplots in R (e.g. biplot with prcomp).

Based on two examples with original data from HMGU problems of presentation and interpretation of biplots are discussed and some solutions are presented in the following. First of all data of soil parameter are analyzed in relation to two experimental factors (with a small number of parameters and a relative high number of objects), secondly OTUs are explored with objects classified in three groups (a high number of OTUs collected for a small number of objects). To avoid subject specific interpretations names of variables and factors are neutralized (results are published). Test versions of R-programs to create modified biplots are available.

2 Multivariate analysis of soil parameter

Data of example 1 are summarized to a matrix with 236 lines (related to different probes) and 13 columns (related to different soil parameters). Besides a general data description structural differences in soil parameters with respect to two experimental factors should be analyzed: one factor with levels A, B, C and another factor with levels 0 and 1.

2.1 unsupervised method / PCA

Biplots for a principal component anlysis can be interpreted as a projection of a p-dimensional space (with p = number of variables) on a two-dimensional plane. Each suchlike projection can be represented as a linear combination of *all* variables. With PCA an *optimal* direction of projection is searched by optimizing the declared variance for each component sucessively. The variance in data can result from desired variance (e.g. based on differences between groups) or undesired variance (e.g. random fluctuation or variance caused by an outlier).

2.1.1 standard biplot based on prcomp and biplot in R

A multivariate data description is performed based on a principal component analysis. Because of the different units of measurements the PCA is build on the correlation matrix. Denominating the 263×13 data matrix with *soil*, labelling the columns with the names of the soil parameters (here soil1 to soil13) and the rows with the combination of factor levels (here A/0, A/1 , B/0, B/1, C/0, C/1) the following R commands produce a standard biplot:

pca1 ← prcomp(soil, scale=TRUE)

biplot(pca1,cex=0.65)

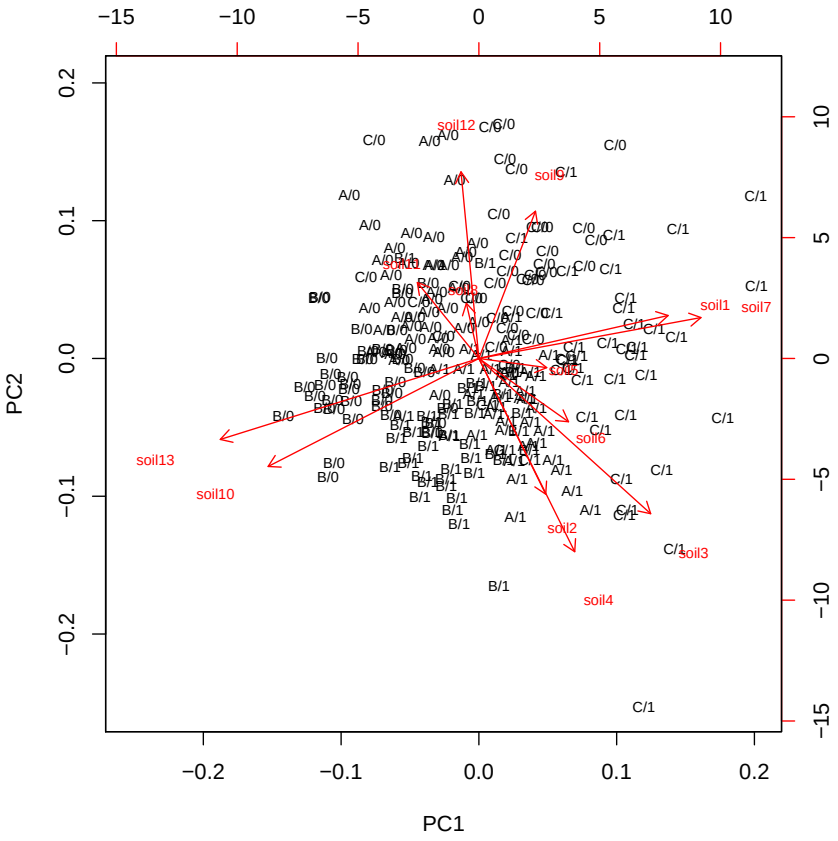

Figure 1: standard biplot based on prcomp and biplot in R

The standard biplot (Figure 1) includes some of the above mentioned *faults*. In particular the *separate canonical scores for samples and variables* (for samples with black tickmarks, for variables with red tickmarks) cannnot be interpreted reasonable.

2.1.2 modified PCA-biplot

Figure 2 shows a modified biplot as the result of one version of the R-program `calpca`. Whereas the main information is unchanged some modifications can be seen:

- Centering of PC axes; aspect ratio = 1
- no scales on PC axes
- additional information about declared variance for each PC

Because samples are labelled with the levels of the experimental factors sample cluster in the biplot which coincide with groups defined by the factors can indivate an association between variables and factors - the information of the experimental design are not used for calculating the principal components.

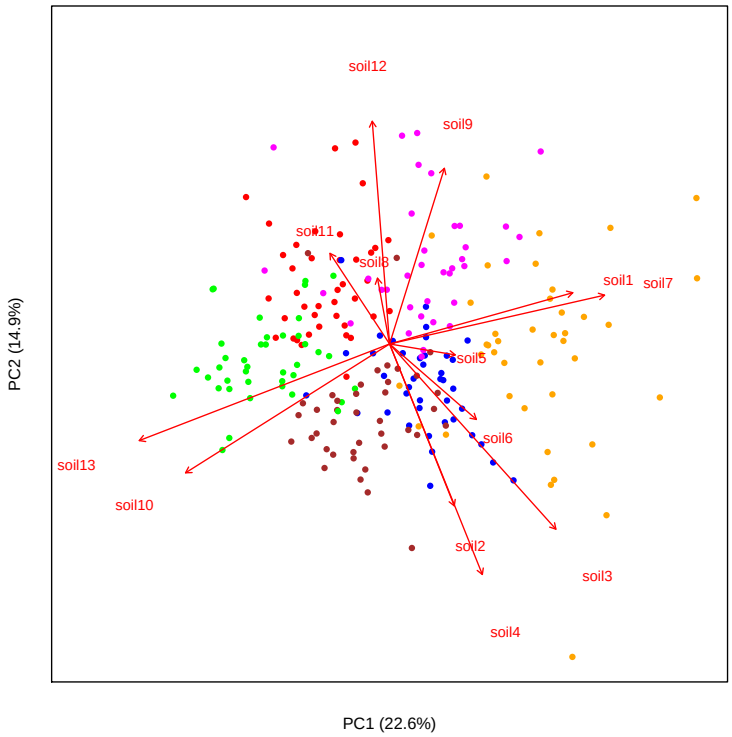

Figure 2: modified PCA-biplot

Even with this modified biplot the interpretetion is restrained by the unusual form of variable axes. The *one-standard error lengths of vectors* and *no scales for the original variables* discourages an interpretation of the results with regard to biological or ecological aspects. Without any doubt the length of variable vectors in classical PCA biplots carry an important information. With only two components the projected data are approximations with different quality of approximation for the different variables. This quality can be measured by the correlation coefficient of approximated and original data. It can be shown that this correlation cefficient is direct proportional to the vector length in the biplot (Table 1).

However a better interpretation of the biplot calls for unresticted axes and especially *scales for the original variables*. There are some utilities for calibration of biplot axes in R: **BiplotGUI** *provides a graphical user interface for the*

	length	r2
soil13	0.54	0.91
soil3	0.50	0.78
soil4	0.50	0.73
soil10	0.49	0.80
soil7	0.45	0.76
soil12	0.45	0.63
soil1	0.39	0.65
soil9	0.37	0.53
soil2	0.35	0.51
soil6	0.23	0.37
soil11	0.22	0.33
soil8	0.14	0.19
soil5	0.13	0.23

Table 1: relation between vector length in biplot and correlation between approximated and original data

construction of, interaction with and manipulation of biplots in R. The package `calibrate` can be used to calibrate axes in scatterplots and biplots in R. Some other existing biplot software can also be used to construct calibrated biplots (**GGEbiplot**, **Genstat**).

In the following we use the R-program `calipca`. Because of clarity calibration of biplot axes should be restricted to some axes. It is advisible to calibrate axes of good approximated variables only. But beyond that also biological or ecological appects can be relevant. Figure 3 shows the example biplot with two calibrated axes for the parameters soil4 (NO_3BS [μgg^{-1})] and soil10 (C_{Total} [%]).

It can be seen that there are some tendency of these soil parameters with respect to the factors: increasing values for NO_3BS (soil10) according to levels C, A, B of factor 1 and higher values of C_{Total} (soil4) in level 1 of factor 2.

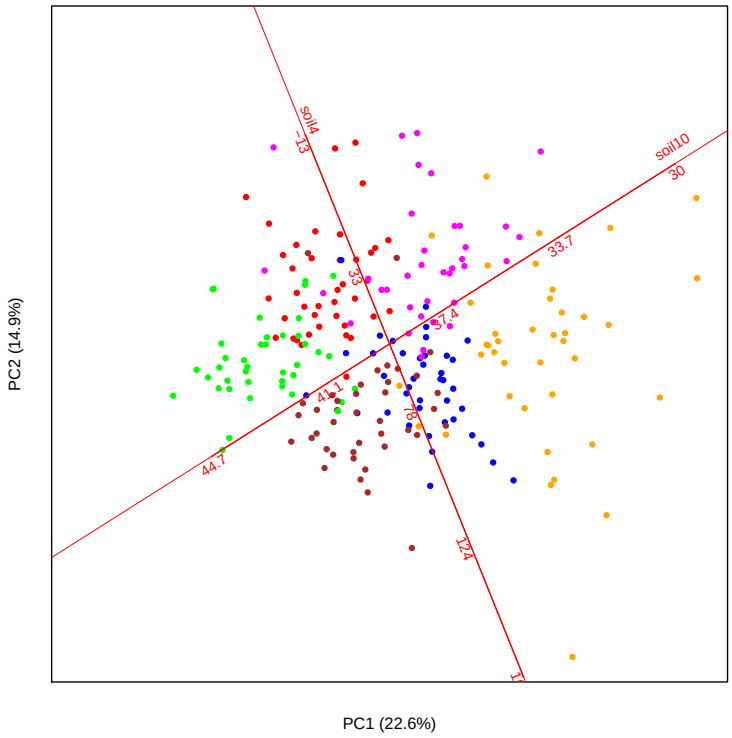

Figure 3: modified PCA-biplot (two axes calibrated)

2.2 supervised methods

With supervised method the *supervisor* information is used to construct the *optimal* direction of projection. While the geometric interpretation of biplots (PCA) - as projection of a higher dimensional space on a two dimensional plane - can give some hints about an association between variables and factors, this interpretation may be misleading using a *supervised* method. Because the *optimal* linear combination of all variables is constructed such that the quotient of between group variance and within group variance is maximized, an visual separation of specified groups is possible, even if there is no significant relation between variables and groups. Therefore all visualizations of results based on e.g. between group analysis should be preceeded by a multivariate test.

2.2.1 multivariate test

To perform a multivariate test a permutational analysis of variance using distance matrices can be used (`adonis` in package `vegan`). The formula is similar

to a normal ANOVA but with a distance matrix on the left side; results have an analog meaning (Table 2).

Performing a between class analysis a randomization test on the between-groups inertia percentage (`randtest(bca)` in package `ade4`) can be added (see chapter 2.2.2).

Under the header *variable selection* often p univariate tests are used with a multiple test correction (see chapter 2.2.3).

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
group2f	5.000	933.817	186.763	20.251	0.306	0.001
Residuals	230.000	2121.183	9.223		0.694	
Total	235.000	3055.000			1.000	

Table 2: ANOVA table (`adonis`)

The permutational analysis of variance shows a significant group effect. Supervised methods can be used to find some additional information about the relation between variables and factors,

2.2.2 `bca` from ade4 package

Figure 4 shows the standard output of `bca` for example 1. A lot of information is summarized combining six figures. Having in mind that the used method - a *diagonal discriminat analysis* - is equivalent to a normal PCA with the group means of all variables, all PCA-biplot informations can be extracted from this figures. The upper left figure (`Canonical weights`) corresponds to the variables part, the upper right figure (`Scores and classes`) to the sample part of the biplot. The figure down left illustrates the declared variances. But this declared variance is only related to the **between groups** variance. That means for instance that with three groups the sum of declared variance for two principal components always equals 100 percent.

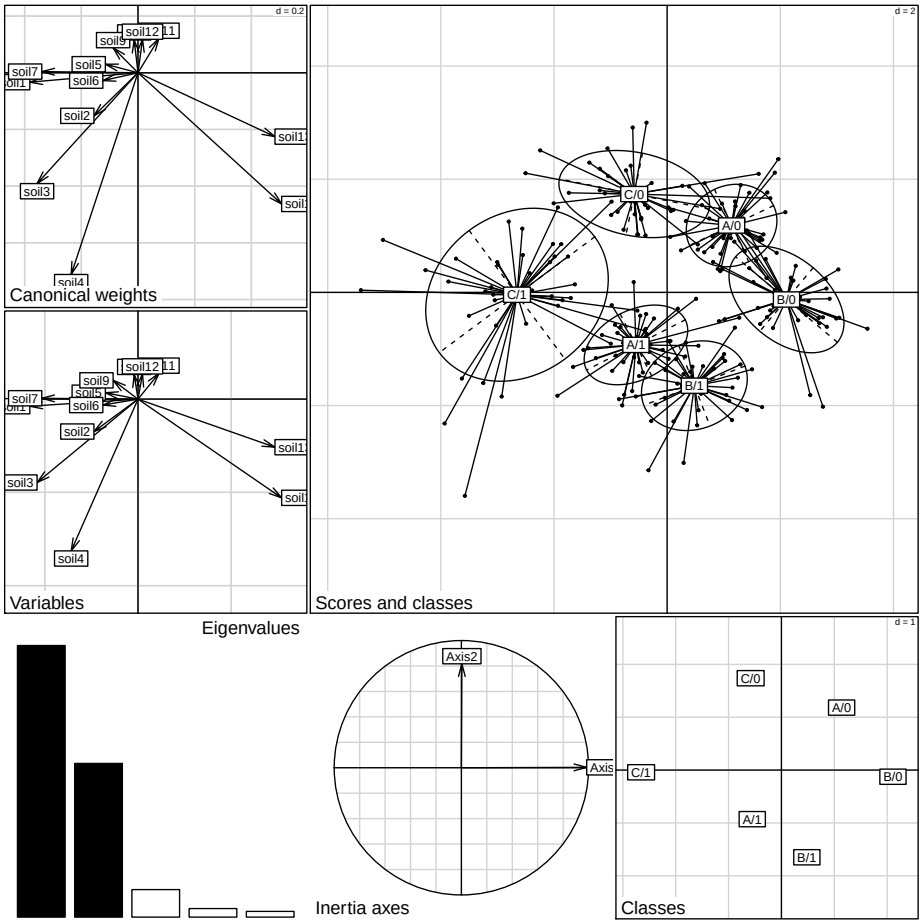

Figure 4: `bca` output

An additional information can be added by running a permutation test on the between-groups inertia percentage. For the data of example 1 the result is:

p-value = 0.001.

2.2.3 modified DDA-biplot

Using the R-program `calidda` we can construct the biplot related to the output of `bca` and essentially based on a PCA of group means (Figure 5).

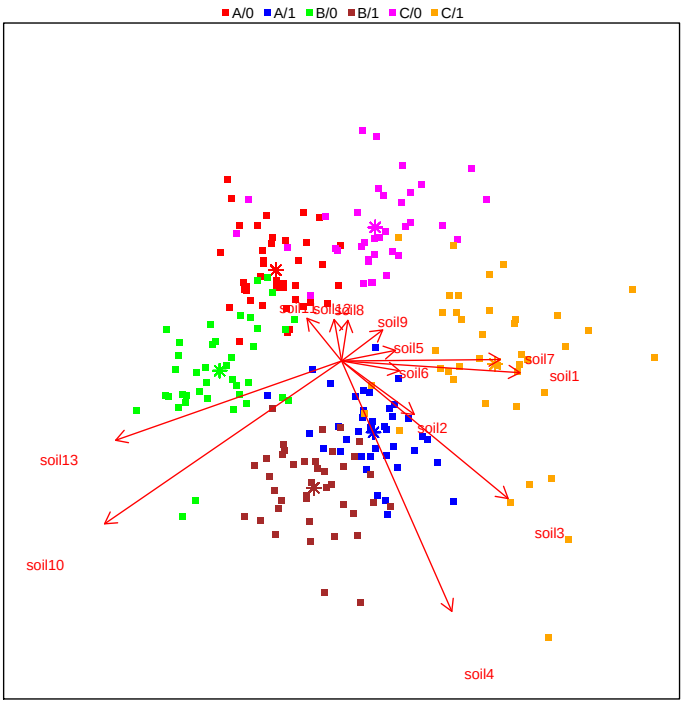

Figure 5: modified DDA-biplot

Again a better interpretation of the biplot calls for calibrated axes. And again the information of vector length in the standard biplot can be used for variable selection. In DDA-biplot this vector length is related to the F-statistic of a univariate ANOVA test for each variable. Accounting for the problem of multiple testing adjusted p-values (*holm* method) produce a list of seven significant variables ($p < 0.01$). This list coincides with the list of the seven longest length biplot vectors (Table 3).

Using the two variables with highest F-value for calibration we can construct the modified DDA-biplot (Figure 6). The dashed lines reflect the projections of the group mean on the two axes and are therefore interpretable as (approximative) values of the two variables.

	length	F.value	adj.pval
soil10	0.930	332.430	0.000
soil4	0.890	190.170	0.000
soil13	0.780	79.720	0.000
soil3	0.700	46.930	0.000
soil1	0.580	29.750	0.000
soil7	0.520	20.270	0.000
soil2	0.290	10.340	0.000
soil6	0.190	2.080	0.210
soil5	0.180	3.510	0.030
soil11	0.180	2.400	0.150
soil9	0.170	2.670	0.110
soil12	0.140	1.020	0.490
soil8	0.130	1.350	0.490

Table 3: relation between vector length in biplot, F value, and adjusted p-value

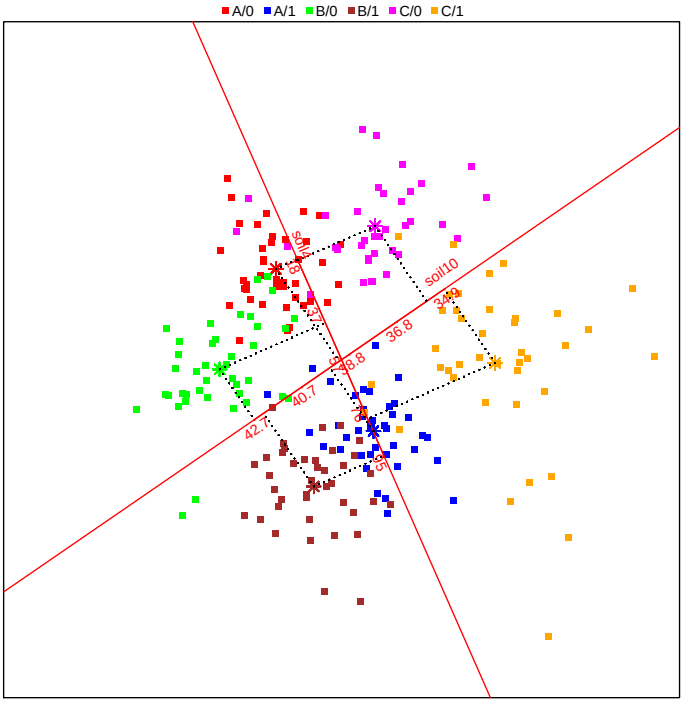

Figure 6: modified DDA-biplot (two axes calbrated)

2.2.4 canonical variate analysis

The *diagonal discriminant analysis* can be seen as a specific form of a canonical variate analysis. The crucial point of this method is the estimation of the covariance or correlation matrix. With p variables p×(p+1)/2 parameters have to be estimated in a *linear discriminant analysis* (LDA). If the numbers of variables is high relatively to the number of samples a robust method for parameter estimation should be used. The *shrinkage* method uses a perturbed estimator of covariance matrix depending on a parameter between 0 and 1. 1 (greatest shrinkage) is related to DDA wheras 0 (no shrinkage) is related to LDA. For data of example 1 LDA is generally applicable. Figure 7 and 8 shows the modified DDA-biplot, Figure 9 and 10 reclassification rate and leave one out crossvalidation rate, which demonstrate a sound group separation.

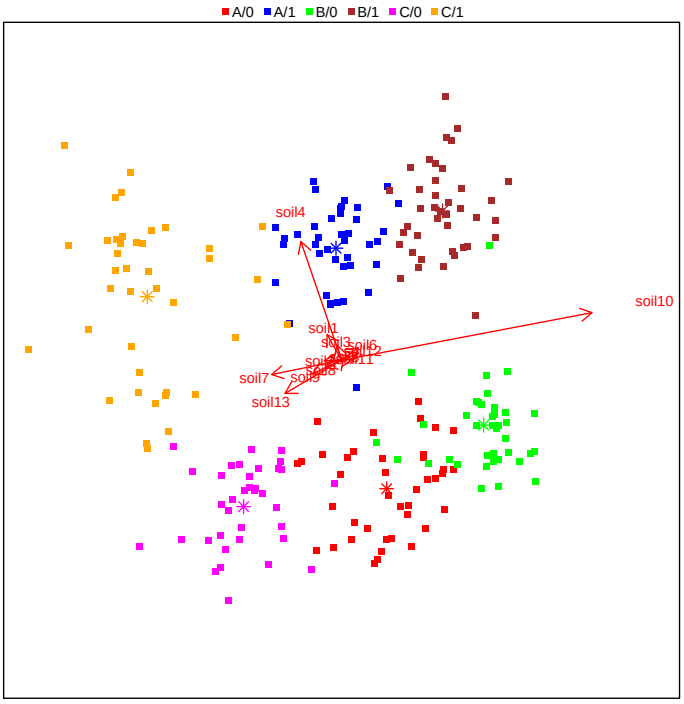

Figure 7: modified LDA-biplot

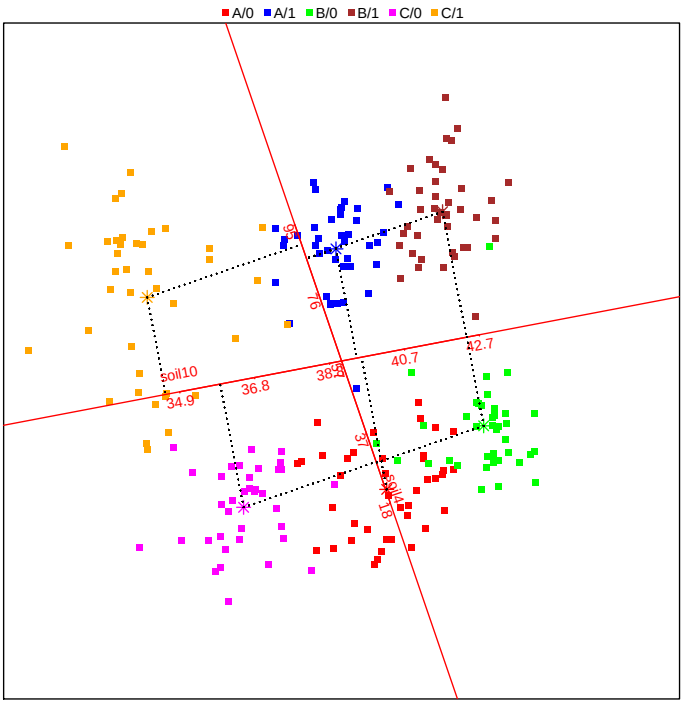

Figure 8: modified LDA-biplot (two axes calibrated)

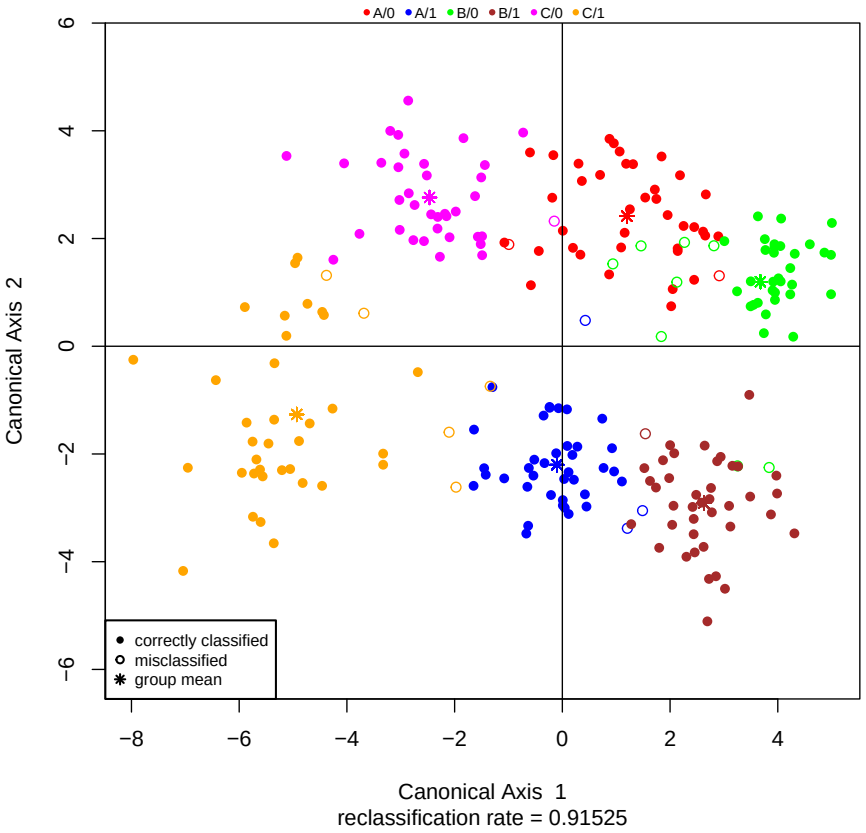

Figure 9: LDA with reclassification

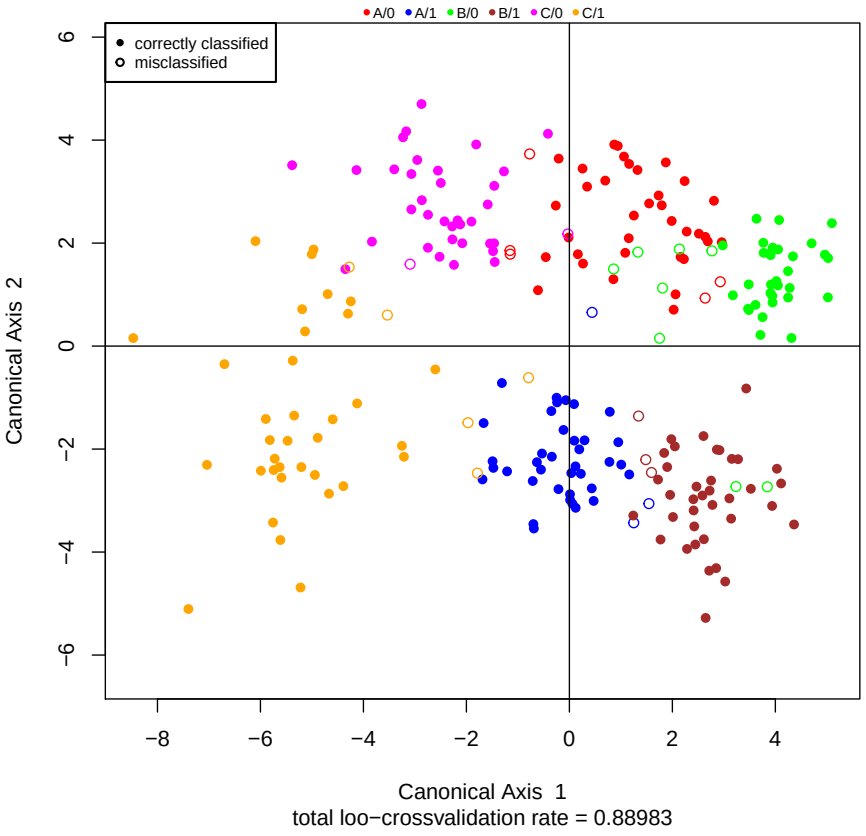

Figure 10: LDA with leave one out cross validation

3 multivariate analysis of OTUs

3.1 unsupervised / PCA

Data of example 2 are combined to a matrix with 25 lines (probes) and 255 columns (absolute abundance of 255 *operational taxonomic units* (OTU). The probes are classified in three groups with levels A (n=7), B and C (n=9 each). Abundance of OTUs is transformed to relativ frequencies only OTUs with at least one probe with relative frequency higer than 1% are considered (84 OTUs).

3.1.1 standard biplot with prcomp and biplot in R

Before application of PCA relative frequencies are sqrt-transformed (*Hellinger*-transformation). With relative frequencies both covariance and correlation based PCA can be used. We applied correlation based PCA. The standard biplot is shown in Figure 11.

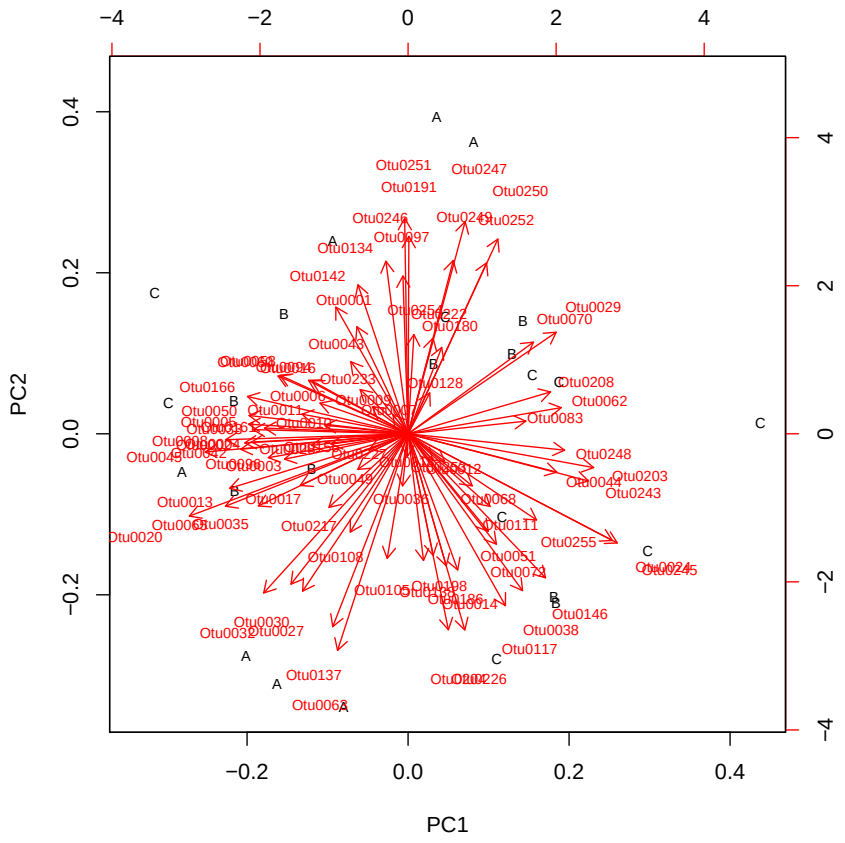

Figure 11: standard biplot based on prcomp and biplot in R

3.1.2 modified biplot

Without variable selection the modified PCA-biplot is like the standard biplot not clearly arranged (Figure 12).

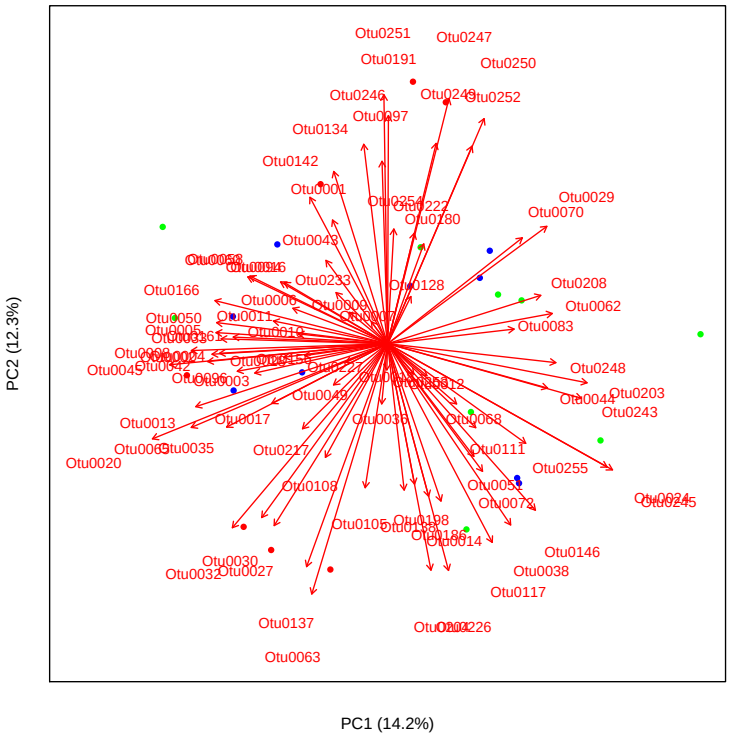

Figure 12: modified PCA-biplot

Modifying the PCA-biplot by choosing two axes for calibration reveals no sample clusters related to the group structure (Figure 13).

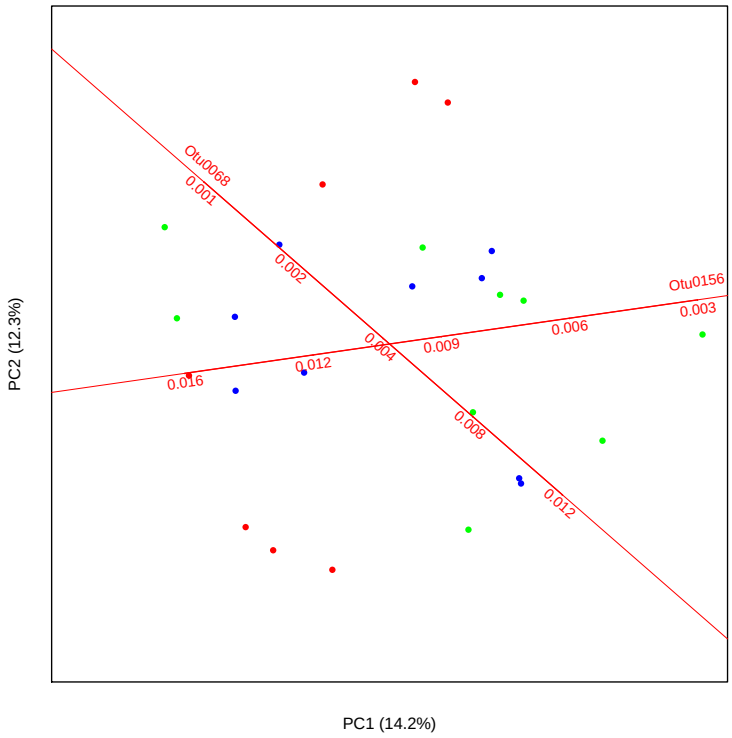

Figure 13: modified PCA-biplot (two axes calbrated)

3.2 supervised Methoden

In any case multivarite test should preceed a supervised analysis.

3.2.1 multivariate test

The following ANOVA table results from a permutational analysis of variance using a distance matrix (Table 4).

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
group1f	2.00	204.29	102.14	1.24	0.10	0.10
Residuals	22.00	1811.71	82.35		0.90	
Total	24.00	2016.00			1.00	

Table 4: ANOVA table (`adonis`)

There is no significant effect. In spite of this fact in the next chapter a between class analysis is performed.

3.2.2 `bca` from ade4 package

Figure 14 shows the standard output of `bca`. The figure may be misleading in case of deduction of a group separation based on the OTUs. It has to be taken into account that the projection direction was constructed using the group information and therefore some form of separation may occur even with *random* data.

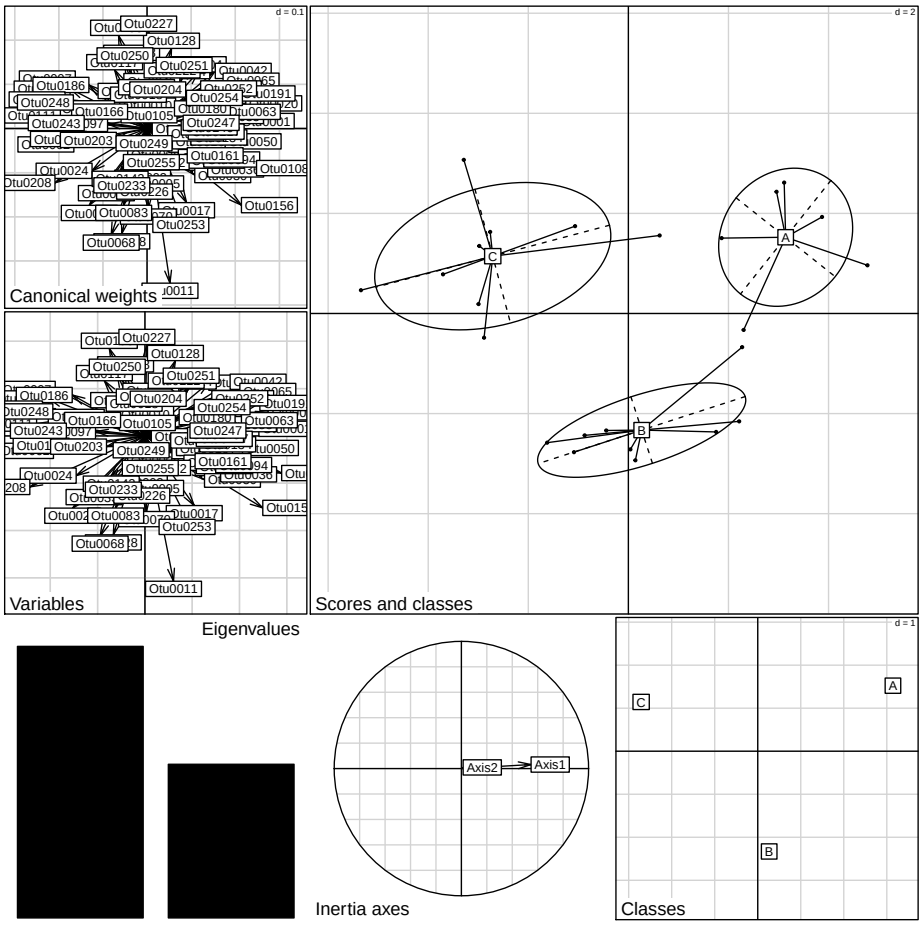

Figure 14: `bca` output

The result for the permutation test on the between-groups inertia percentage is:

p-value = 0.114.

Again no significant effect can be seen.

3.2.3 modified DDA-biplot

Again it should be mentioned that the modified DDA-biplot can also be used for variable selection. Ranking of variables with respect to the length of variables is equivalent to a ranking to univariate test statistics (ANOVA F-statistic) as done e.g. by the `Metastats` program in the **mothur** project. This equivalence is

absolutely valid if the number of groups equals two or three and approximately correct otherwise.

Figure 15 shows the modified DDA-biplot with the ten best-ranked varaibles.

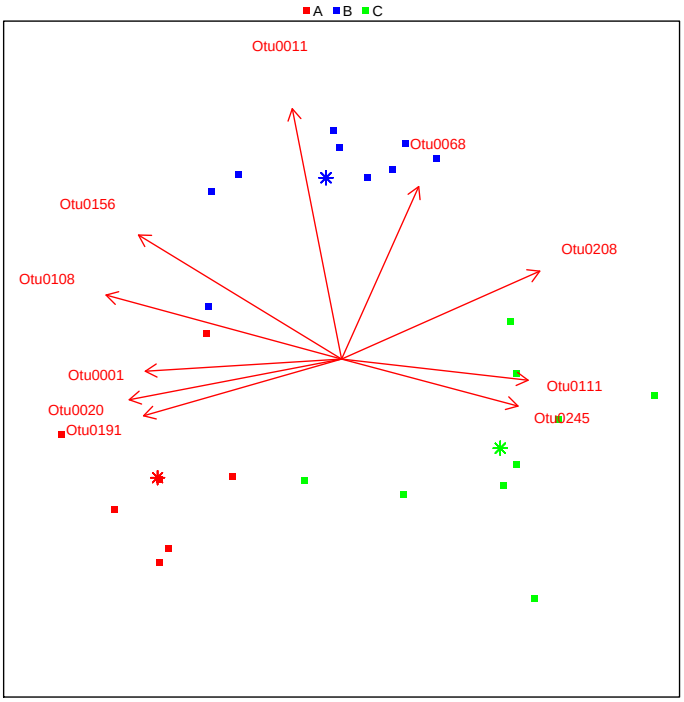

Figure 15: modified DDA-biplot with variable selection

Table 5 presents some additional information about these ten variables; to handle the problem of multiple testing one column is added with adjusted p-values (holm-method). The list with variables with adj.pval < 0.05 would be empty.

	length	F.value	adj.pval
Otu0011	0.626	7.105	0.350
Otu0108	0.599	6.165	0.621
Otu0156	0.584	5.700	0.830
Otu0208	0.532	4.349	1.000
Otu0020	0.531	4.317	1.000
Otu0191	0.506	3.780	1.000
Otu0001	0.483	3.344	1.000
Otu0068	0.464	3.015	1.000
Otu0111	0.461	2.967	1.000
Otu0245	0.448	2.766	1.000

Table 5: variable selection based on ANOVA F-value or vector-length in biplot

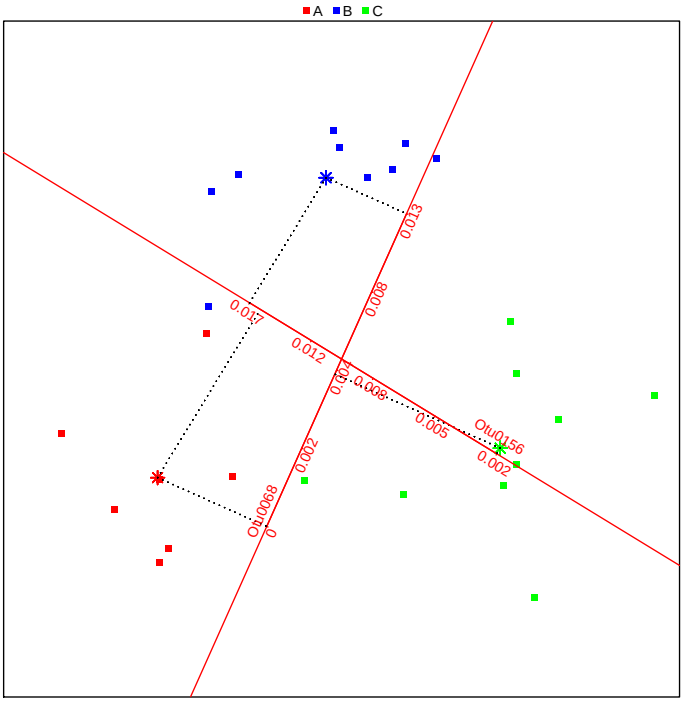

Figure 16: modified DDA-biplot (two axes calibrated)

The calibration for two axes in Figure 16 is related to the relative abundances of OTUs despite the sqrt- transformation used running the PCA.

Figure 17 and 18 shows reclassification rate and leave one out cross validation

rate, the latter demonstrating that no sound group separation is possible.

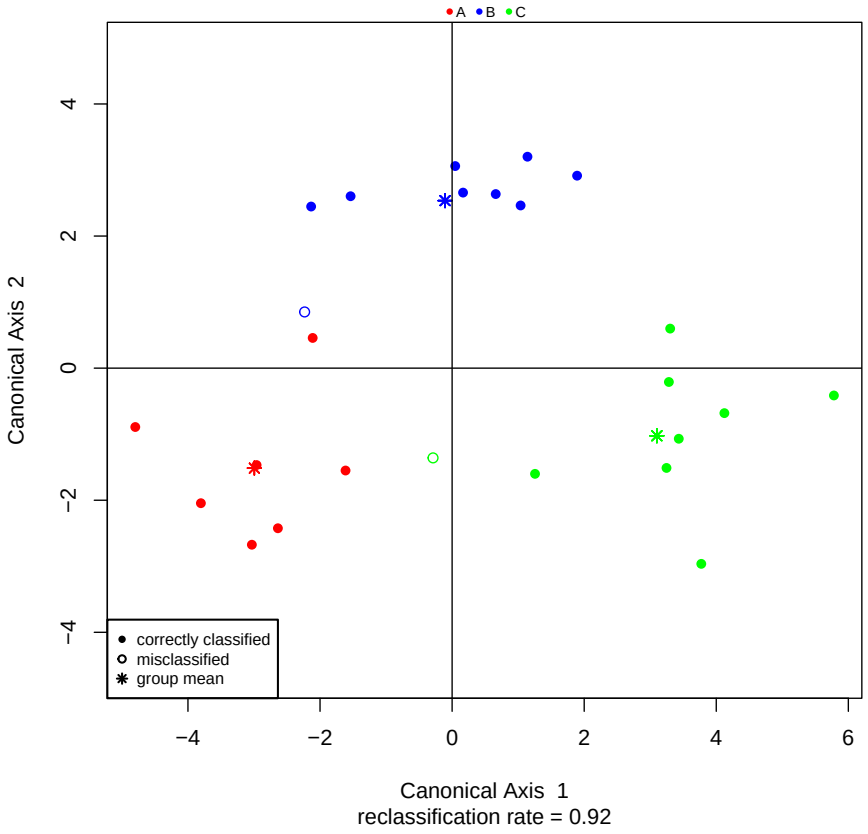

Figure 17: DDA reclassification

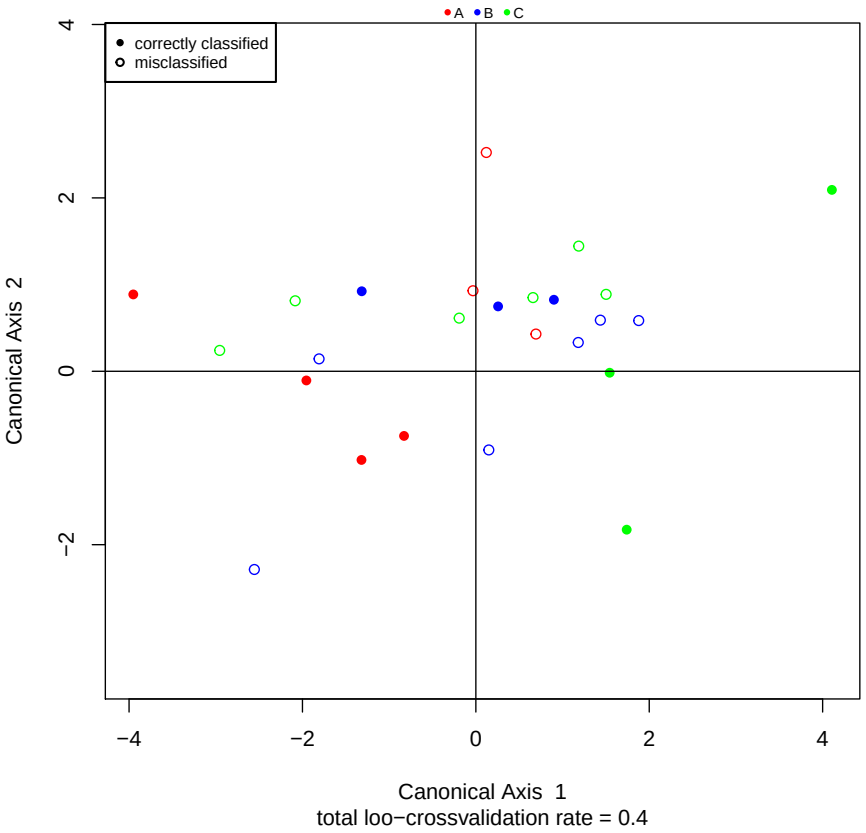

Figure 18: DDA loo cross validation